

Critical behavior and universality classes for an algorithmic phase transition in sparse reconstruction

Mohammad Ramezanali*

*Department of Physics and Astronomy, Rutgers University,
136 Frelinghuysen Rd, Piscataway, NJ 08854 USA*

Partha P. Mitra†

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11734 USA

Anirvan M. Sengupta‡

*Department of Physics and Astronomy, Rutgers University,
136 Frelinghuysen Rd, Piscataway, NJ 08854 USA and
Center of Quantitative Biology, Rutgers University,
176 Frelinghuysen Rd, Piscataway, NJ 08854 USA*

(Dated: November 19, 2015)

Optimization problems with sparsity-inducing penalties exhibit sharp algorithmic phase transitions when the numbers of variables tend to infinity, between regimes of ‘good’ and ‘bad’ performance. The nature of these phase transitions and associated universality classes remain incompletely understood. We analyze the mean field equations from the cavity method for two sparse reconstruction algorithms, Basis Pursuit and Elastic Net. These algorithms are associated with penalized regression problems $(\mathbf{y} - \mathbf{H}\mathbf{x})^2 + \lambda V(\mathbf{x})$, where $V(\mathbf{x})$ is an algorithm-dependent penalty function. Here λ is the penalty parameter, \mathbf{x} is the N -dimensional, K -sparse solution to be recovered, \mathbf{y} is the M dimensional vector of measurements, and \mathbf{H} is a known ‘random’ matrix. In the limit $\lambda \rightarrow 0$ and $N \rightarrow \infty$, keeping $\rho = K/N$ fixed, exact recovery is possible for sufficiently large values of fractional measurement number $\alpha = M/N$, with an algorithmic phase transition occurring at a known critical value $\alpha_c = \alpha(\rho)$. We find that Basis Pursuit $V(\mathbf{x}) = \|\mathbf{x}\|_1$ and Elastic Net $V(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \frac{\lambda_2}{2} \|\mathbf{x}\|_2^2$ belong to two different universality classes. The Mean Squared Error goes to zero as $MSE \sim (\alpha_c - \alpha)^2$ as α approaches α_c from below for both algorithms. However, for Basis Pursuit, precisely on the phase transition line $\alpha = \alpha_c$, $MSE \sim \lambda^{4/3}$ for Basis Pursuit, whereas $MSE \sim \lambda$ for Elastic Net. We also find that in presence of additive noise, within the perfect reconstruction phase $\alpha > \alpha_c$ there is a non-zero setting for λ that minimizes MSE, whereas at $\alpha = \alpha_c$ a finite λ always increases the MSE.

I. INTRODUCTION

Optimization problems with sparsity penalties have become widely popular. One example is that of compressed sensing, where priori structure in the signals in the form of sparsity in a suitable domain is exploited to reduce the number of measurements required to retrieve the signal. Early work in this area by Candès and Donoho [3, 5] exploited a combination of a convex relaxation of a strict convexity cost using an ℓ_1 -penalty, together with a random choice of measurement matrices to define the problem. A striking feature of this work was the computation of an algorithmic phase transition boundary separating a ‘good’ regime in which perfect reconstruction is possible in suitable limits, from a ‘bad’ regime where such reconstruction is impossible.

A typical statement of the sparse retrieval problem is an ill-posed linear equation, $\mathbf{y} = \mathbf{H}\mathbf{x}$ (noise free case), where \mathbf{y} is an M dimensional measurement vector, \mathbf{H} is an $M \times N$ known measurement matrix, and \mathbf{x} is an N dimensional unknown parameter vector ($M < N$). Assume that \mathbf{y} is generated by $\mathbf{H}\mathbf{x}_0$, where \mathbf{x}_0 is the N dimensional vector to be retrieved from the knowledge of \mathbf{y} and \mathbf{H} . It is a priori known that \mathbf{x}_0 has at most K nonzero components. The task is to reconstruct this unknown vector. The ill-posedness of the underdetermined linear system is removed by imposing a sparsity constraint. For typical \mathbf{H} , as long as the number of unknowns, K , is less than the number of measurements M , these linear equalities have a unique sparse solution with high probability. A common formulation for the problem is to pose it as an optimization problem, defining $\hat{\mathbf{x}}(\lambda\sigma^2) = \arg \min_{\mathbf{x}} \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{H}\mathbf{x})^2 + \lambda V$. In this sparse

*Electronic address: mrr@physics.rutgers.edu

†Electronic address: mitra@cshl.edu

‡Electronic address: anirvans@physics.rutgers.edu

estimation framework, the purpose of the cost function V is to penalize the number of nonzero entries of \mathbf{x} so that the sparsity property of the source is carried over to the solution $\hat{\mathbf{x}}$. The so-called ℓ_0 norm, $\|\mathbf{x}\|_0 = \lim_{p \rightarrow 0+} \|\mathbf{x}\|_p$, where $\|\mathbf{x}\|_p = \sqrt[p]{\sum_a |x_a|^p}$, really counts the number of nonzero elements of \mathbf{x} .

This is a non-convex optimization problem and it has exponential computational complexity. However, it was Chen et al [4] who introduced Basis Pursuit technique that uses the ℓ_1 - norm for evaluating sparsity as a computationally tractable convex relaxation of this optimization problem. They showed that the quest for the sparsest solution could be tackled as a convex programming task, often leading to the correct result for sufficiently small K (sufficient sparsity). Another approach is to combine ℓ_1 and ℓ_2 norms, i.e. $V(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \frac{\lambda_2}{2} \|\mathbf{x}\|_2^2$ which is known as Elastic Net [16]. ℓ_1 and ℓ_2 penalized estimation methods both shrink the estimates of the regression coefficients towards zero to prevent overfit due to either collinearity of the covariates or high-dimensionality. Although both methods are shrinkage methods, the effects of ℓ_1 and ℓ_2 penalization are quite different in practice. Applying an ℓ_2 penalty tends to result in all small but non-zero regression coefficients, whereas applying an ℓ_1 penalty tends to result in many regression coefficients shrunk exactly to zero and a few other regression coefficients with comparatively little shrinkage. Combining ℓ_1 and ℓ_2 penalties tends to give a result in between, with fewer regression coefficients set to zero than in a pure ℓ_1 setting, and more shrinkage of the other coefficients. The amount of shrinkage is determined by tuning parameters, λ_1 and λ_2 . It was shown by Zou and Hasties [16] that Elastic Net is effective at grouping highly correlated variables, i.e. they are either selected or removed from the model as a group.

In particular, for measurement matrices that have independent and identically distributed (iid) Gaussian entries, it is shown that Basis Pursuit requires as low as $M > O(K \log(N/K))$ measurements for perfect reconstruction [3, 5]. In several papers, analyses based on the message-passing method, and the replica formalism borrowed from statistical physics, indicate that the performance failure of the ℓ_1 norm minimization method and other analogous algorithms with polynomial time complexity occurs at a sharp boundary as $N \rightarrow \infty$, with $\frac{M}{N}$ and $\frac{K}{N}$ being held fixed, analogous to a second-order (continuous) phase transition [6, 8, 9]. This is an algorithmic phase transition or zero-one law, where recovery fails or succeeds with high probability, jumping from zero to one at the transition boundary. In this paper, we explore these transition boundaries using an alternative analytical approach based on the cavity method which we presented in our earlier work [13]. Second order transitions are classified by their critical exponents, which characterize the behavior at the critical point. We obtain these critical exponents in various cases, for example with additive-noise and non-zero trade-off parameter. Eventually, we gain additional insights into the problem as well as insights about a more general set of optimization problems.

II. NOTATION AND PAPER OUTLINE

For matrices, we use boldface capital letters like \mathbf{H} , and we use \mathbf{H}^T , $\text{tr}(\mathbf{H})$, to denote the transpose and trace, respectively. For vectors, we use boldface small letters like \mathbf{x} with x_a representing the a^{th} element of \mathbf{x} . We use $[\dots]_{\text{vars}}^{\text{av}}$ to denote quenched averages, with the relevant quenched variables indicated in the subscript. In particular, this average depends on two random variables \mathbf{x}_0 and \mathbf{H} that are drawn from distribution $P_0(\mathbf{x}_0)$ and $\mathcal{P}(\mathbf{H})$. For a Gaussian random variable x with mean μ and variance v , we write the pdf as $\mathcal{N}(x; \mu, v)$ and, for the special case of $\mathcal{N}(x; 0, 1)$, we abbreviate the pdf as $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and write the cdf as $\Phi(x) = \int_x^\infty dz \phi(z)$. Dirac's delta function is written as $\delta(x)$ and δ_{mn} is the Kronecker delta symbol.

The rest of the paper is organized as follows. The problem formulation and brief recapitulation of the cavity method presented previously in [13] is given in Section III. We first treat the simple case of Ridge Regression [15]. Then, since the ℓ_1 -regularization is practically the most studied special case of the general setup, we apply the cavity approach and find a simple way to arrive at the two phases and the phase boundary in an insightful way and recover the known analytical formulation of results for Basis Pursuit [4]. Next, we study this phase transition boundary in various cases of additive-noise and non-zero trade-off parameter and find the behavior of the error as a function of these parameters and their scaling exponents at the phase boundary. Finally, we extend our analysis to the Elastic Net [16] and obtain new ones. Conclusions and Summary are provided in Section VII.

III. PROBLEM FORMULATION AND METHODS

Consider the standard compressed sensing (CS) setup, $\mathbf{y} = \mathbf{H}\mathbf{x}_0 + \boldsymbol{\zeta}$, where it is assumed that \mathbf{H} in $\Re^{M \times N}$ represents the measurement ($M \leq N$) sensing system, the sparse vector \mathbf{x}_0 in \Re^N is unknown, and the vector $\boldsymbol{\zeta}$ is the measurement error with $E[\boldsymbol{\zeta}] = 0$, $E[\boldsymbol{\zeta}\boldsymbol{\zeta}'] = \sigma_\zeta^2 \mathbf{I}_M$. One possible way to reconstruct \mathbf{x} from \mathbf{y} , given the measurement

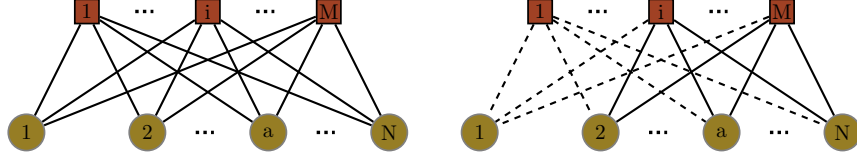


FIG. 1: top) A system of variable nodes (circles) and constraint nodes (squares) and bottom) a cavity surrounding a single node ‘a’ and a single constraint ‘i’, that have been removed consecutively via two cavity steps from the system shown by the dashed lines.

matrix \mathbf{H} is to minimize the following penalized regression function with the estimated vector $\hat{\mathbf{x}}$ defined by

$$\hat{\mathbf{x}}(\vartheta) = \arg \min_{\mathbf{x}} \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{H}\mathbf{x})^2 + \lambda V(\mathbf{x}). \quad (1)$$

$\vartheta = \lambda\sigma^2$ is a non-negative parameter giving relative weight between the first and second term in Eq. (1) and $V : \Re^N \rightarrow \Re$ a fixed non-negative regularization function also known as the cost function. Here like most cases in the penalized regression problem, we will focus on a $V(\mathbf{x}) = \sum_a U(x_a)$ that is convex and separable. In the sparse estimation framework, a well studied case is sparsity promoting regularizing function $U(x) = \lambda|x|$. It is known that for this cost function, Eq. (1) gives an exact reconstruction of \mathbf{x}_0 in a certain region of parameter space. Using $\mathbf{y} = \mathbf{H}\mathbf{x}_0$, we write the minimization of Eq. (1) as an optimization over the function $\mathcal{E}(\mathbf{u})$ in terms of the error variable $\mathbf{u} = \mathbf{x} - \mathbf{x}_0$.

$$\mathcal{E}(\mathbf{u}) = \frac{(\mathbf{H}\mathbf{x}_0 - \mathbf{H}\mathbf{x} - \boldsymbol{\zeta})^2}{2\sigma^2} + V(\mathbf{x}) = \frac{1}{2\sigma^2} (\mathbf{H}\mathbf{u} + \boldsymbol{\zeta})^2 + V(\mathbf{u} + \mathbf{x}_0) \quad (2)$$

Notice that the new cost function $\mathcal{E}(\mathbf{u})$ itself is now a function of the input signal \mathbf{x}_0 . This mainly reflects the fact that we are interested in studying the statistical behavior of its minima over the distribution of instances of \mathbf{x}_0 and \mathbf{H} rather than the traditional optimization of $\mathcal{E}(\mathbf{u})$ for a given \mathbf{x}_0 and \mathbf{H} . The distribution of estimation error may be quantified by a suitable norm, for example $f(\mathbf{x}, \mathbf{x}_0) = \frac{1}{N} (\mathbf{x} - \mathbf{x}_0)^2 = \frac{1}{N} \mathbf{u}^2$, as a measure of the inaccuracy of the reconstruction. The average of this quantity corresponds to the mean squared estimation error (MSE).

$$\text{MSE} \equiv \frac{1}{N} [(\mathbf{x} - \mathbf{x}_0)^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} = \frac{1}{N} [\mathbf{u}^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} \quad (3)$$

Although in general \mathbf{H} could be drawn from a non-Gaussian distribution, at this point we consider the special case in which $\mathbf{H}^T \mathbf{H}$ is nearly proportional to a unit matrix, i.e.

$$[H_{ia}]^{\text{av}} = 0 \quad (4)$$

$$[H_{ia} H_{jb}]^{\text{av}} = \frac{1}{M} \delta_{ij} \delta_{ab} \quad (5)$$

And the vector \mathbf{x}_0 is a random sample drawn from a distribution $P_0(\mathbf{x}_0) = \prod_a p_0(x_{a0})$. Here we consider the sparsity promoting distribution $p_0(x_{a0})$ which has a continuous part and a delta function at origin:

$$p_0(x_a) = \rho \pi(x_a) + (1 - \rho) \delta(x_a). \quad (6)$$

In the case of a full random measurement matrix, we adapt the cavity method in two steps to derive self-consistency conditions on MSE when the regularization term $V(\mathbf{x})$ is such that a unique solution exists. The cavity mean field equations are named after a physical context in spin systems in solid state physics [11, 12]. The goal is to take into account the non-trivial dependencies by estimating the reaction of all the other ‘spin’ variables when a single spin is removed from the system, thereby leaving a ‘cavity’ (see Fig. 1). This leads to a considerable simplification by utilizing the fact that the system of variables are fully connected and the so-called local susceptibility matrix χ known to physicists plays a key role in the system [13]. In particular, in the asymptotic limit of large M and N , certain quantities (e.g. MSE and average local susceptibility, $\bar{\chi}(\mathbf{x})$) converge, i.e. becomes independent of the detailed realization of the matrix \mathbf{H} in the large N and M limits. In this limit, the optimization of Eq. (2) breaks down into a collection of effectively independent optimization as the followings:

TABLE I: Symbols

Symbols	
Symbol	Description
u_a	Measure of residual error $x_a - x_{0a}$
q	Mean squared error (MSE)
α	Measure for the number of constraints, $\frac{M}{N}$
ρ	Measure for the sparsity, $\frac{K}{N}$
λ_1	ℓ_1 -norm regression coefficient
λ_2	ℓ_2 -norm regression coefficient
σ^2	Error variance on the constraint $\mathbf{y} = \mathbf{H}\mathbf{x}$
ϑ	$\lambda\sigma^2$
σ_{eff}^2	Effective σ^2 given in the asymptotic limit of large M, N
θ	$\lambda\sigma_{\text{eff}}^2$
σ_ξ^2	$\frac{q}{\alpha}$
σ_ζ^2	Variance of external noise
τ	$\frac{\theta}{\sigma_\xi}$

Table that presents symbols used in this article.

Proposition 1 (Effective Individual Optimization).

$$\hat{u}_a = \min_{u_a} \left\{ \frac{1}{2\sigma_{\text{eff}}^2} (u_a^2 - 2\xi_a u_a) + U(u_a + x_{0a}) \right\} \quad (7)$$

$$\xi_a \in \mathcal{N}(\xi; 0, \sigma_\xi^2) \text{ with } \sigma_\xi^2 \equiv \sigma_\zeta^2 + \frac{q}{\alpha} \quad (8)$$

$$q \equiv \sum_a [\hat{u}_a^2]_{x_0, \xi}^{\text{av}} \quad (9)$$

$$\sigma_{\text{eff}}^2 \equiv \sigma^2 + \frac{\bar{\chi}}{\alpha} \quad (10)$$

$$\bar{\chi} \equiv \frac{1}{N} \sum_a \chi^{aa} \quad (11)$$

The quantity q is the sum of the squared of error residuals, i.e. MSE. In addition, local susceptibility is obtained via $\hat{u}_a(f) - \hat{u}_a(0) = \chi^{aa} f_a$ with $f_a \rightarrow 0$ and $\hat{u}_a(f)$ is carried out by minimizing $\min_{u_a} \left\{ \frac{1}{2\sigma_{\text{eff}}^2} (u_a^2 - 2\xi_a u_a) + U(u_a + x_{0a}) - f_a u_a \right\}$. In the end, summing over χ^{aa} 's for all the instances of measurement matrix and then taking average over all nodes yields to the average local susceptibility, $\bar{\chi}$, and thus σ_{eff}^2 . As it is presented in [13], the asymptotic estimates of the local susceptibilities is given by $[\chi^{aa}(\mathbf{x})]_{\text{av}} = \left[U''(\hat{u}_a + x_{0a}) \delta_{ab} + \frac{1}{\sigma_{\text{eff}}^2} \right]^{-1}$.

An equivalent formulation for the above has been obtained by [8, 9] with the replica approach. In order to make a connection with these studies, one should replace $\bar{\chi}$ by the quantity $\beta\Delta Q$ in Eq. (10) in which¹²

$$\Delta Q \equiv [\langle (u - \langle u \rangle)^2 \rangle]_{x_0, \xi}^{\text{av}}. \quad (13)$$

However, the writers believe introducing β and taking limits $\beta \rightarrow \infty$ leads to a less transparent result in the study of phase transition and is unnecessary.

¹ Here β is a spurious dimensionless quantity playing the role of inverse temperature and $\langle \dots \rangle$ is the average over thermal fluctuations of \mathbf{u} with ensemble probability function P_{eff}

$$P_{\text{eff}}(u | x_0, \xi) = \frac{1}{Z(x_0, \xi)} e^{-\beta \mathcal{E}_{\text{eff}}(u; x_0, \xi)}. \quad (12)$$

² One can notice that the quantity ΔQ is nothing but the thermal fluctuations in \mathbf{u} and $\beta\Delta Q$ is in fact can be identified as a local susceptibility due to the fluctuation-dissipation theorem [10].

As we will see in the next section, we solve the effective individual optimization in Proposition 1 for the penalty function of the form $\lambda|x|^q$ with $q = 1, 2$. We point out the importance of $\bar{\chi}$ over $\beta\Delta Q$ to distinguish phases around the zero-temperature transition described by Donoho and Tanner [7].

To facilitate such adoption, we summarize the symbols used in the next section in the Table III.

IV. MAIN RESULT

A. Ridge Regression

We start by considering the simplest form of regularization with $U(x) = \frac{\lambda}{2}x^2$, a penalty function that does not impose sparsity on the solutions. This is just a noise-free ridge regression with Tikhonov regularization [15]

$$\hat{\mathbf{x}}(\vartheta = \lambda\sigma^2) = \arg \min_{\mathbf{x}} \left\{ \frac{1}{2\sigma^2} (\mathbf{H}(\mathbf{x} - \mathbf{x}_0))^2 + \frac{\lambda}{2} \mathbf{x}^2 \right\}. \quad (14)$$

We could explicitly minimize \mathbf{x} and proceed with our analysis using random matrix theory; however, we will apply first the self-consistency formalism we have developed (Proposition 1).

$$\min_u \left\{ \frac{1}{2\sigma_{\text{eff}}^2} (u^2 - 2\xi u) + \frac{\lambda}{2} (u + x_0)^2 \right\} \quad (15)$$

Recalling that $u = x - x_0$ and identifying $\theta = \lambda\sigma_{\text{eff}}^2$, minimization of Eq. (15) gives

$$\hat{x} = \frac{x_0 + \xi}{1 + \theta} \quad (16)$$

This result can be used to determine σ_{ξ}^2 in Eq. (8)

$$\sigma_{\xi}^2 = \frac{q}{\alpha} = \frac{1}{\alpha} [u^2]_{x_0, \xi}^{\text{av}} = \frac{\sigma_{\xi}^2 + \theta^2 \rho [x_0^2]_{x_0}^{\text{av}}}{\alpha(1 + \theta)^2}. \quad (17)$$

where $[\dots]_{x_0}^{\text{av}}$ means average over $\pi(x_0)$. One can see that with the ridge regression penalty function in 1, local susceptibility is the same everywhere:

$$\bar{\chi} = \left[\lambda + \frac{1}{\sigma^2 + \frac{\bar{\chi}}{\alpha}} \right]^{-1} \implies \theta = \left(\frac{1}{\lambda\bar{\chi}} - 1 \right)^{-1}. \quad (18)$$

In particular in the $\vartheta \rightarrow 0$ limit, i.e. the minimal ℓ_2 norm subject to linear constraints $\mathbf{H}\mathbf{x} = \mathbf{H}\mathbf{x}_0$, with $\lambda\bar{\chi} = 1 - \alpha$ gives $\theta = \alpha^{-1} - 1$. With the knowledge of θ , the Eqs. (16), and (17) lead us to

$$\hat{x} = \alpha(x_0 + \xi) \quad (19)$$

$$\sigma_{\xi}^2 = \frac{(1 - \alpha)\rho}{\alpha} [x_0^2]_{x_0}^{\text{av}}. \quad (20)$$

which is the same conclusion from a formal singular value decomposition point of view (see Appendix A).

Remark 1. The estimated x can be seen as a Gaussian variable, with αx_0 as its mean and $(1 - \alpha)\alpha\rho[x_0^2]_{x_0}^{\text{av}}$ as its variance. When $x_0 = 0$, we expect the fluctuation of x around zero to be of the order $((1 - \alpha)\alpha\rho[x_0^2]_{x_0}^{\text{av}})^{1/2}$. We could have a simple threshold θ so that if $|x| < \theta$ we set x_0 to zero. We could compute the false positive and false negative rates of such a procedure. When $\rho \ll \alpha/(1 - \alpha)$, it is possible to choose a threshold θ such that $((1 - \alpha)\alpha\rho[x_0^2]_{x_0}^{\text{av}})^{1/2} \ll \theta \ll \alpha([x_0^2]_{x_0}^{\text{av}})^{1/2}$. With such a threshold, both error rates would be small.

B. Basis Pursuit: ℓ_1 -norm Minimization

In this section, we reconsider the much-analyzed case where the penalty function is the ℓ_1 norm of \mathbf{x} [3, 5, 7]. The reconstructed sparse solution is given by

$$\hat{\mathbf{x}}(\vartheta) = \min_{\mathbf{x}} \left\{ \frac{1}{2\sigma^2} (\mathbf{H}(\mathbf{x} - \mathbf{x}_0))^2 + \lambda \|\mathbf{x}\|_1 \right\}. \quad (21)$$

Like in the case of ridge regression, we aim to solve the equations in proposition 1 for the potential $U(x) = \lambda|x|$ self-consistently. To determine θ , once again we look at the local susceptibilities in 1. In this case $U''(x)$ is zero everywhere except at $x = 0$, where it is formally infinite. Consequently,

$$\begin{aligned}\chi^{aa} &= 0, \text{ if } x_a = 0 \\ \chi^{aa} &= \sigma_{\text{eff}}^2, \text{ otherwise.}\end{aligned}\tag{22}$$

We define $\hat{\rho}$ to be the estimated sparsity, i.e. the fraction of x_a 's that are non-zero. Therefore $\bar{\chi} = \hat{\rho}\sigma_{\text{eff}}^2$ ($\lambda\bar{\chi} = \hat{\rho}\theta$) and

$$\sigma_{\text{eff}}^2 = \sigma^2 + \frac{\bar{\chi}}{\alpha} = \sigma^2 + \frac{\hat{\rho}\sigma_{\text{eff}}^2}{\alpha}\tag{23}$$

implying

$$\theta(1 - \frac{\hat{\rho}}{\alpha}) = \vartheta\tag{24}$$

Remark 2. The equation $\theta(1 - \frac{\hat{\rho}}{\alpha}) = \vartheta$ is central to understanding the $\vartheta \rightarrow 0$ limit and the associated phase transition. When ϑ goes to zero, we either have $\theta = 0$ ($\hat{\rho} \neq \alpha$) or $\hat{\rho} = \alpha$ ($\theta \neq 0$). These two conditions correspond to the two phases of the system, the first being the perfect reconstruction phase and the second, the non-zero error regime. In terms of average local susceptibility, the first phase has $\bar{\chi} = \hat{\rho}\theta = 0$, while the second one has $\bar{\chi} \neq 0$.

Now we can set up the notation for the single variable optimization problem to find the value for σ_ξ^2 (\propto MSE) in these two regimes. More precisely, by searching for the solutions to

$$\min_u \left\{ \frac{1}{2\sigma_{\text{eff}}^2} (u^2 - 2\xi u) + \lambda|u + x_0| \right\}\tag{25}$$

we arrive at the following soft-thresholding function for the estimated value of \hat{x} that we will denote by $\eta_{\text{soft}}(t; \theta)$, with the variable $t = x_0 + \xi$.

Definition 1 (Soft Thresholding Function).

$$\eta_{\text{soft}}(t; \theta) = \begin{cases} t - \theta & \text{if } \theta \leq t, \\ 0 & \text{if } -\theta \leq t \leq \theta, \\ t + \theta & \text{if } t < -\theta. \end{cases}\tag{26}$$

According to remark 2, the perfect reconstruction regime which ends to the phase boundary from above is the case where, as ϑ becomes small, θ becomes small as well. From Eq. (26), there are three sources of error that can contribute to σ_ξ^2 in this regime (illustrated in Fig. 2):

a) ($x_0 \neq 0 \rightarrow \hat{x} = 0$)

Here x_0 was initially non-zero, but the estimated \hat{x} , due to the shift by ξ , has fallen into the $[-\theta, \theta]$ interval and then been truncated to zero. One can see that since θ is small, the probability of this event can be ignored for the time being³.

b) ($x_0 \neq 0 \rightarrow \hat{x} \neq x_0$)

For non-zero x_0 that does not get set to zero, the contribution to MSE is

$$\rho[(\hat{x} - x_0)^2]_{x_0, \xi}^{\text{av}} = \rho[(\xi - \theta \text{sgn}(\hat{x}))^2]_{x_0, \xi}^{\text{av}} = \rho(\sigma_\xi^2 + \theta^2)\tag{27}$$

³ Under this circumstance, if ξ remains of order one, then the error is dominated by ξ , i.e. $q(\text{MSE}) = \sigma_\xi^2$. However, this is not consistent with $\sigma_\xi^2 = q/\alpha$, unless $\sigma_\xi^2 = 0$. Hence in this regime, we need to consider a σ_ξ^2 that is comparable to θ . Therefore, as $\vartheta \rightarrow 0$, we will have $\sigma_\xi^2 \rightarrow 0$ and $q \rightarrow 0$, making the reconstruction perfect, i.e. the limit when $\vartheta, \theta, \sigma_\xi^2 \rightarrow 0$ with $\frac{\theta}{\sigma_\xi}$ of order one.

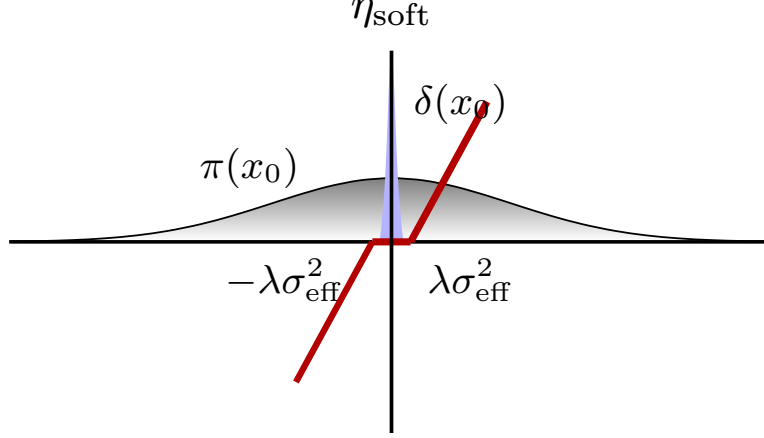


FIG. 2: The soft thresholding function (in red) defined in (26). The non-zero entries of the sparse vector \mathbf{x}_0 drawn from random distribution is represented by π (in grey) and the zero components are represented by delta function (in blue) (6)

c) ($x_0 = 0 \rightarrow \hat{x} \neq 0$)

Another source of error is the event when the x_0 is zero but \hat{x} has fallen outside the interval $[-\theta, \theta]$ and has been estimated to be non-zero. In this case, the contribution to MSE is

$$\begin{aligned} (1 - \rho)[\hat{x}^2]_{x_0, \xi}^{\text{av}} &= 2(1 - \rho) \int_{\theta}^{\infty} d\xi \frac{1}{\sqrt{2\pi\sigma_{\xi}^2}} e^{-\frac{\xi^2}{2\sigma_{\xi}^2}} (\xi - \theta)^2 \\ &= 2\sigma_{\xi}^2(1 - \rho) \{ (1 + \tau^2)\Phi(\tau) - \tau\phi(\tau) \} \end{aligned} \quad (28)$$

with $\tau = \frac{\theta}{\sigma_{\xi}}$. Adding up these contributions from Eq. (27) and (28), we get the total MSE, q (i.e. $\alpha\sigma_{\xi}^2$). Therefore using Eq. (8), $\sigma_{\xi}^2 = q/\alpha$, and the knowledge of $\theta = 0$ lead to the first parametric expression for the perfect reconstruction phase:

$$\alpha = 2(1 - \rho) \{ (1 + \tau^2)\Phi(\tau) - \tau\phi(\tau) \} + \rho(1 + \tau^2). \quad (29)$$

To determine $\hat{\rho}$, one can notice that if $x_0 = 0$, we have to have $|\xi| > \theta$ to lead to a non-zero x . On the other hand, since θ is small, a non-zero x_0 remains non-zero with probability approaching one. Counting all sources of the non-zero \hat{x} 's, then we have ⁴

$$\hat{\rho} = 2(1 - \rho)\Phi(\tau) + \rho. \quad (30)$$

Recall that in the error-prone phase $\hat{\rho} = \alpha$ (Remark 2). This is due to the fact that q, σ_{ξ}^2 and therefore θ need to be non-zero in this regime. If the transition happens continuously, the condition for the phase boundary is $\alpha = \hat{\rho} = 2(1 - \rho)\Phi(\tau) + \rho$. Hence the relation between α and ρ at the phase boundary is obtained by solving and eliminating τ from

$$\alpha = 2(1 - \rho) \{ (1 + \tau^2)\Phi(\tau) - \tau\phi(\tau) \} + \rho(1 + \tau^2) \quad (31)$$

$$\alpha = 2(1 - \rho)\Phi(\tau) + \rho \quad (32)$$

⁴ Note that $\hat{\rho} > \rho$, even in the perfect reconstruction phase. That is because a fraction of x_a 's remain non-zero as long as $\vartheta > 0$, and vanish only in the $\vartheta \rightarrow 0$ limit.

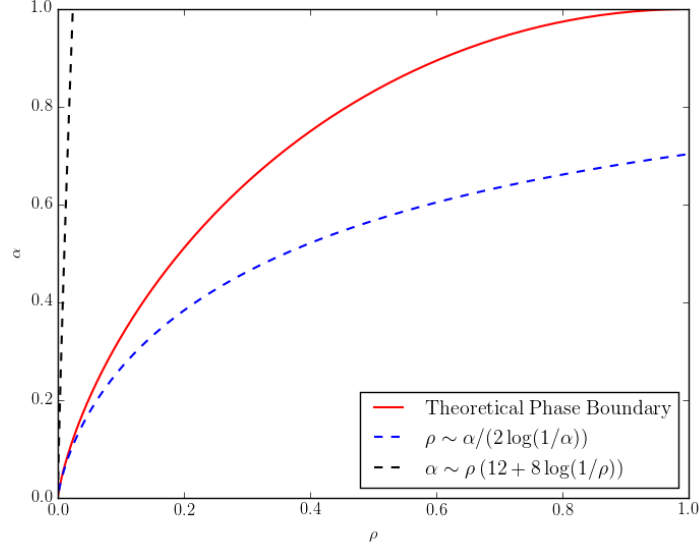


FIG. 3: The red curve is the theoretical phase boundary obtained by solving Eq. (31) and (32). As $\rho \rightarrow 0$ this boundary is of the form $\rho = \alpha / (2 \log(1/\alpha))$ as it is shown by dashed blue curve. The black dashed curve shows one of the restricted isometry property bounds [14]. Perfect recovery occurs above the red curve.

Alternatively, Eq. (31) and (32) can be solved for α and ρ at the phase boundary and expressed parametrically as a function of τ :

$$\alpha = \frac{2\phi(\tau)}{\tau + 2(\phi(\tau) - \tau\Phi(\tau))} \quad (33)$$

$$\rho/\alpha = 1 - \frac{\tau\Phi(\tau)}{\phi(\tau)} \quad (34)$$

This leads to the phase diagram showing the transition from absolute success to absolute failure depicted in Fig. 3.

Remark 3. In the extremely sparse limit, $\rho \ll 1$, one can obtain a more explicit asymptotic relation between α and ρ . In this limit τ is large, and the dominant contributions are the second term, $\rho(1 + \tau^2)$, from Eq. (31) and the first term, $2(1 - \rho)\Phi(\tau)$, from Eq. (32). Consequently,

$$\begin{aligned} \alpha &\approx \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{\tau^2}{2}}}{\tau}, \quad \rho \approx \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{\tau^2}{2}}}{\tau^3} \\ \implies \rho &\approx \frac{\alpha}{\tau^2} \approx \frac{\alpha}{2 \log(1/\alpha)} \end{aligned} \quad (35)$$

Therefore, in sparse limit, we have $\rho \sim \alpha / (2 \log \frac{1}{\alpha})$ (see Fig. 3). Apart from a coefficient, this result has similar form to the bounds from the restricted isometry property [2].

V. CRITICAL EXPONENTS

To get a better understanding of the nature of this phase transition and characterizing its behavior as one decreases α from above $\alpha_c(\rho)$ to below, we should search for solutions of Eq. (36) and (37) in the error-prone regime where both θ and σ_ξ^2 remain $O(1)$. In this case, we have to deal carefully with the possibility that \hat{x} has been set to zero, because

$x_0 + \xi$ fell within $\pm\theta$. It is straightforward to show that the self-consistency equation for σ_ξ^2 becomes

$$\begin{aligned} \alpha = & \alpha \frac{\sigma_\xi^2}{\sigma_\xi^2} + 2(1 - \rho) \{ (1 + \tau^2) \Phi(\tau) - \tau \phi(\tau) \} \\ & + \rho \left[\tau_0^2 \{ 1 - \Phi(\tau + \tau_0) - \Phi(\tau - \tau_0) \} \right. \\ & + (1 + \tau^2) \{ \Phi(\tau + \tau_0) + \Phi(\tau - \tau_0) \} \\ & \left. - (\tau - \tau_0) \phi(\tau + \tau_0) - (\tau + \tau_0) \phi(\tau - \tau_0) \right]_{x_0}^{\text{av}} \end{aligned} \quad (36)$$

where $[\dots]_{x_0}^{\text{av}}$ means average over $\pi(x_0)$ and $\tau_0 = \frac{x_0}{\sigma_\xi}$. The quantity τ and functions $\Phi(\tau)$ and $\phi(\tau)$ are defined as before. In addition, the parametric expression of Eq. (32) becomes

$$\alpha = \frac{\vartheta}{\theta} + 2(1 - \rho) \Phi(\tau) + \rho \left[\Phi(\tau + \tau_0) + \Phi(\tau - \tau_0) \right]_{x_0}^{\text{av}} \quad (37)$$

One should notice that in Eqs. (36), (37) we included extra terms $\alpha \frac{\sigma_\xi^2}{\sigma_\xi^2}$ coming from the additive noise and $\frac{\vartheta}{\theta}$ from not setting ϑ to zero, respectively.

In order to better understand the behavior close to the transition where θ and σ_ξ^2 are small, we rewrite Eqs. (36) and (37) as ⁵

$$\alpha = \alpha \frac{\sigma_\xi^2}{\sigma_\xi^2} + A_2(\rho, \tau) - \rho [\psi_\xi(\tau_0, \tau)]_{x_0}^{\text{av}} \quad (38)$$

$$\alpha = \alpha \frac{\vartheta}{\theta} + A_0(\rho, \tau) - \rho [\psi_\theta(\tau_0, \tau)]_{x_0}^{\text{av}} \quad (39)$$

where $\psi_\xi(\tau_0, \tau)$, $\psi_\theta(\tau_0, \tau)$ are even functions of τ_0 that falls off quickly as τ_0 becomes much larger than 1. For convenience we also defined

$$\begin{aligned} A_2(\rho, \tau) &= 2(1 - \rho) \{ (1 + \tau^2) \Phi(\tau) - \tau \phi(\tau) \} + \rho(1 + \tau^2) \\ A_0(\rho, \tau) &= 2(1 - \rho) \Phi(\tau) + \rho \end{aligned} \quad (40)$$

In Fig. 4, the behavior of these two functions are shown for a specific ρ .

Remark 4. The transition boundary is where these two curves intersect at the point τ_c . Note that $\frac{dA_2}{d\tau} = 2 \frac{A_2 - A_0}{\tau}$. Thus, at the transition point τ_c , $\frac{dA_2}{d\tau} = 0$, i.e. A_2 behaves like $\sim \delta\tau^2$ (A_0 goes as $\sim -\delta\tau$). As we will see in section VD, this relation will not be valid for Elastic Net. Therefore, we expect to have different critical behavior near the transition point for Elastic Net than Basis Pursuit.

Moreover, we can write

$$\begin{aligned} [\psi_\xi(\tau_0, \tau)]_{x_0}^{\text{av}} &= \int dx_0 \pi(x_0) \psi_\xi\left(\frac{x_0}{\sigma_\xi}, \tau\right) \\ &= \sigma_\xi \int d\tau_0 \pi(\sigma_\xi \tau_0) \psi_\xi(\tau_0, \tau) \end{aligned} \quad (41)$$

and get the same expression for $\psi_\theta(\tau_0, \tau)$. Thus, the small σ_ξ^2 behavior of these averages depends on how $\pi(x)$ behaves at small x . When $\pi(x) \sim Fx^\gamma$ with $\gamma > -1$:

$$[\psi_\xi(\tau_0, \tau)]_{x_0}^{\text{av}} \approx \sigma_\xi^{\gamma+1} \int d\tau_0 \tau_0^\gamma \psi_\xi(\tau_0, \tau) \sim \sigma_\xi^{\gamma+1} \quad (42)$$

⁵ Note that, when $|\tau_0| = \frac{|x_0|}{\sigma_\xi} \rightarrow \infty$, $\Phi(\tau + \tau_0) + \Phi(\tau - \tau_0) \rightarrow 1$ and $(\tau - \tau_0)\phi(\tau + \tau_0), (\tau + \tau_0)\phi(\tau - \tau_0) \rightarrow 0$. The τ_0 dependent expression inside $[\dots]_{x_0}^{\text{av}}$ in Eq. (36) goes from $2\{(1 + \tau^2)\Phi(\tau) - \tau\phi(\tau)\}$ to $1 + \tau^2$ as τ_0 goes from zero to infinity. We wrote this expression as $1 + \tau^2 - \psi_\xi(\tau_0, \tau)$.

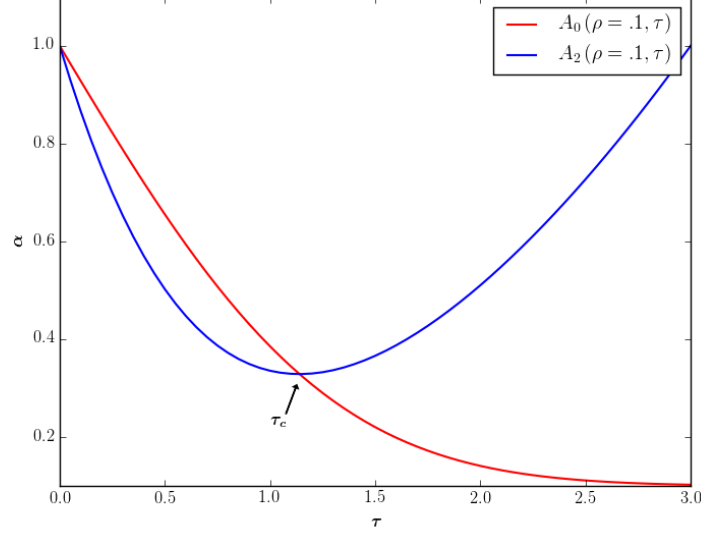


FIG. 4: The transition boundary is where the red and blue curves meet at the critical τ_c .

Similarly $[\psi_\theta(\tau_0, \tau)]_{x_0}^{\text{av}} \sim \sigma_\xi^{\gamma+1}$. Thus, the perturbations added to phase boundary Eqs. (31) and (32) are of the order of $\sigma_\xi^{\gamma+1}$. Accordingly, in the case of a gapped distribution so that $\pi(x) = 0$ when $|x| < \Delta$, we have:

$$[\psi_\xi(\tau_0, \tau)]_{x_0}^{\text{av}} \approx \sigma_\xi \int_{\Delta} d\tau_0 \psi_\xi(\tau_0, \tau) \sim e^{-\frac{\Delta^2}{\sigma_\xi^2}} \sigma_\xi \quad (43)$$

And $[\psi_\theta(\tau_0, \tau)]_{x_0}^{\text{av}} \sim e^{-\frac{\Delta^2}{\sigma_\xi^2}} \sigma_\xi$

A. Into the Error-prone regime ($\vartheta \rightarrow 0$ & $\sigma_\xi^2 = 0$)

To find an estimate for the mean-squared error by entering into the error-prone regime, we express the phase boundary as $\alpha = \alpha_c(\rho), \tau = \tau_c(\rho)$ by solving Eqs. (31), and (32). To explore close to the phase boundary, we can write $\alpha = \alpha_c(\rho) - \delta\alpha$ and $\tau = \tau_c(\rho) - \delta\tau$. Since the perturbations to Eqs. (31), (32) for the case of $\pi(x) \sim Fx^\gamma$ are of the order $\sigma_\xi^{\gamma+1}$, from equation (38) we get

$$\delta\alpha \sim \sigma_\xi^{\gamma+1} = \left(\frac{q}{\alpha}\right)^{\frac{\gamma+1}{2}} \quad (44)$$

Therefore, for nonzero terms drawn from a distribution with nonzero density at the origin, Eq. (44) tells us that the mean square error rises as

$$q(MSE) \sim (\alpha_c - \alpha)^{\frac{2}{\gamma+1}} \quad (45)$$

Similarly, for $\pi(x)$ with a gap, we get a sharp rise for the error:

$$q \sim \frac{1}{\ln(1/(\alpha_c - \alpha))} \quad (46)$$

Remark 5. The additional insight is that although the phase boundary $\alpha_c(\rho)$ does not depend on the distribution of non-zeros, the rise of the error does and becomes sharper when non-zero components are farther from zero. Moreover, the rise is continuous, i.e. it is a second-order phase transition and its critical exponent depends on the behavior of $\pi(x_0)$ near $x_0 = 0$.

B. Role of an Additive Noise ($\vartheta \rightarrow 0$ & $\sigma_\zeta^2 \neq 0$)

To examine the behavior of Eqs. (31), and (32) close to the phase boundary within the presence of noise, once again, we Taylor expand them around the transition point where $\alpha = \alpha_c(\rho)$ and $\tau = \tau_c(\rho)$. Therefore, for the case of $\pi(x) \sim Fx^\gamma$, Eqs. (38), (39) in terms of perturbing variables $\delta\alpha$ and $\delta\tau$ become:

$$\delta\alpha = \alpha_c \frac{\sigma_\zeta^2}{\sigma_\xi^2} + C\delta\tau^2 - D\sigma_\xi^{\gamma+1} + \dots \quad (47)$$

$$\delta\alpha = -C'\delta\tau - D'\sigma_\xi^{\gamma+1} + \dots \quad (48)$$

Where C , D , C' and D' are functions of ρ and τ_c , and ' \dots ' contains higher order corrections. From Eq. (48), we have $\delta\tau = -(D'/C')\sigma_\xi^{\gamma+1}$ which, by substitution to the first Eq. (47), gives

$$0 = \alpha_c \frac{\sigma_\zeta^2}{\sigma_\xi^2} + \frac{C D'^2}{C'^2} \sigma_\xi^{2+2\gamma} - D\sigma_\xi^{\gamma+1} \implies \sigma_\xi^2 \propto (\sigma_\zeta^2)^{2/(3+\gamma)}$$

which we arrived at it by taking into account that $\sigma_\xi^2 \rightarrow 0^+$. With a similar calculation in the case with gapped distribution, we obtain

$$\sigma_\xi^2 \propto \frac{1}{\ln(1/\sigma_\zeta^2)} \quad (49)$$

C. ϑ Trade-off in the Noisy system ($\vartheta \neq 0$ & $\sigma_\zeta^2 \neq 0$)

In the previous subsection, we considered the role of additive Gaussian noise in the behavior of the phase boundary near the transition from perfect reconstruction to the error regime. However, one should take into consideration that in most situations noise arises from several sources and there is no good estimation of either the level or distribution of the noise. Therefore, there is often a trade-off between the least squares of the residual and the ℓ_1 norm of the solution. If the regularization is too much, the regularized solution does not fit the given signal properly as the residual error is too large. If the regularization is too small, the fit will be good but error will be more. One can control this trade-off and the sparsity of the solution by proper selection of the regularization parameter ϑ . In the noise-free case, Taylor expansion of Eqs. (31), and (32) close to the transition point leads to:

$$\delta\alpha = C\delta\tau^2 - D\sigma_\xi^{\gamma+1} + \dots \quad (50)$$

$$\delta\alpha = \alpha_c \frac{\vartheta}{\theta} - C'\delta\tau - D'\sigma_\xi^{\gamma+1} + \dots \quad (51)$$

From Eq. (50), we have $\delta\tau = (D/C)^{1/2}\sigma_\xi^{\frac{\gamma+1}{2}}$ which by substitution into Eq. (51) and by taking into account that $\theta \sim \sigma_\xi$ gives

$$0 = \alpha_c \frac{\vartheta}{\theta} - \frac{C' D^{1/2}}{C^{1/2}} \sigma_\xi^{\frac{\gamma+1}{2}} - D\sigma_\xi^{\gamma+1} \implies \sigma_\xi^2 \propto \vartheta^{\frac{4}{\gamma+3}} \quad (52)$$

Similar calculation with the gapped distribution gives

$$\sigma_\xi^2 \propto \frac{1}{\ln(1/\lambda)} \quad (53)$$

As we mentioned earlier, a more interesting question would be that at what value of ϑ , we will get the minimum error in the presence of noise. By adding noise to the system and expanding Eqs. (39) and (38) in terms of perturbing variables $\delta\alpha$ and $\delta\tau$, we have

$$\delta\alpha = \alpha_c \frac{\sigma_\zeta^2}{\sigma_\xi^2} + C\delta\tau^2 - D\sigma_\xi^{\gamma+1} + \dots \quad (54)$$

$$\delta\alpha = \alpha_c \frac{\vartheta}{\theta} - C'\delta\tau - D'\sigma_\xi^{\gamma+1} + \dots \quad (55)$$

TABLE II: ℓ_1 -norm Minimization

$\pi(x) \sim Fx^\gamma$ with $\gamma > -1$	
Input Variables	Scaling Functions
$\lambda \rightarrow 0, \sigma_\zeta^2 = 0$	$MSE \sim (\alpha_c - \alpha)^{2/(1+\gamma)}$
$\lambda \rightarrow 0, \sigma_\zeta^2 \neq 0$	$MSE \sim (\sigma_\zeta^2)^{2/(3+\gamma)}$
$\lambda \neq 0, \sigma_\zeta^2 = 0$	$MSE \sim \lambda^{4/(3+\gamma)}$
$\lambda \neq 0, \sigma_\zeta^2 \neq 0$	$MSE \sim \lambda^{2/(2+\gamma)}$ $\&\zeta \sim (\sigma_\zeta^2)^{2/(3+\gamma)}$
$\pi(x) = 0$ for $ x < \Delta$	
$\lambda \rightarrow 0, \sigma_\zeta^2 = 0$	$MSE \sim \frac{1}{\ln(1/(\alpha_c - \alpha))}$
$\lambda \rightarrow 0, \sigma_\zeta^2 \neq 0$	$MSE \sim \frac{1}{\ln(1/\sigma_\zeta^2)}$
$\lambda \neq 0, \sigma_\zeta^2 = 0$	$MSE \sim \frac{1}{\ln(1/\lambda)}$
$\lambda \neq 0, \sigma_\zeta^2 \neq 0$	$MSE \sim \frac{1}{\ln(1/\sigma_\zeta^2)}$ $\&\zeta \sim \frac{1}{\ln(1/\lambda)}$

Table presents critical exponents for ℓ_1 -norm minimization near phase transition in terms of relevant input parameters, α , λ , and σ_ζ^2 . In order to relate with compressed-sensing literature, we have set $\sigma^2 = 1$, i.e. $\vartheta = \lambda$.

To have a solution, we get $\sigma_\xi^2 \sim \vartheta^{\frac{2}{\gamma+2}}$ and $\sigma_\xi^2 \sim (\sigma_\zeta^2)^{2/(\gamma+3)}$. Therefore, by tuning ϑ to $(\sigma_\zeta^2)^{\frac{\gamma+2}{\gamma+3}}$, the minimum error occurs. Similarly for the gapped non-zero distribution, $\sigma_\xi^2 \sim \frac{1}{\ln(1/\vartheta)}$ and $\sigma_\xi^2 \sim \frac{1}{\ln(1/\sigma_\zeta^2)}$. These scaling functions and critical exponents are summarized in the table II.

D. Elastic Net

As a quick application of our zero temperature cavity method, we consider how the phase transition would be affected if we generalize our penalty function $V(\mathbf{x})$ by adding a quadratic term $|\mathbf{x}|^2$ to the ℓ_1 norm. This penalty function is used in the Elastic Net method of variable selection and regularization [16]. The optimization problem becomes

$$\hat{\mathbf{x}}_{\text{EN}} = \min_{\mathbf{x}} \left\{ \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{H}\mathbf{x})^2 + \lambda_1 |\mathbf{x}| + \frac{\lambda_2}{2} |\mathbf{x}|^2 \right\} \quad (56)$$

In the noiseless reconstruction problem, $\mathbf{y} = \mathbf{H}\mathbf{x}_0$. We take the limit $\sigma^2 \rightarrow 0$ and choose the distribution of \mathbf{H} and \mathbf{x}_0 to be the same as in the previous sections.

Now $U''(x) = \lambda_2$ everywhere except at $x = 0$, where it is formally infinite, leading to

$$\begin{aligned} \chi^{aa} &= 0, \text{ if } x_a = 0 \\ \chi^{aa} &= \frac{\sigma_{\text{eff}}^2}{1 + \lambda_2 \sigma_{\text{eff}}^2}, \text{ otherwise.} \end{aligned} \quad (57)$$

Once more we define $\hat{\rho}$ to be fraction of x_a s that are non-zero. Then $\bar{\chi} = \frac{\hat{\rho} \sigma_{\text{eff}}^2}{1 + \lambda_2 \sigma_{\text{eff}}^2}$ and

$$\sigma_{\text{eff}}^2 = \sigma^2 + \frac{\bar{\chi}}{\alpha} = \sigma^2 + \frac{\hat{\rho} \sigma_{\text{eff}}^2}{\alpha(1 + \lambda_2 \sigma_{\text{eff}}^2)} \quad (58)$$

implying

$$\sigma_{\text{eff}}^2 \left\{ 1 - \frac{\hat{\rho}}{\alpha(1 + \lambda_2 \sigma_{\text{eff}}^2)} \right\} = \sigma^2 \quad (59)$$

In the $\sigma^2 \rightarrow 0$ limit, the two phases are given by, $\sigma_{\text{eff}}^2 = 0$ or $\hat{\rho} = \alpha(1 + \lambda_2 \sigma_{\text{eff}}^2)$. Again, the perfect reconstruction phase has $\bar{\chi} = \frac{\hat{\rho} \sigma_{\text{eff}}^2}{1 + \lambda_2 \sigma_{\text{eff}}^2} = 0$ and the error-prone regime has $\bar{\chi} = \frac{\hat{\rho} \sigma_{\text{eff}}^2}{1 + \lambda_2 \sigma_{\text{eff}}^2} = \alpha \sigma_{\text{eff}}^2 \neq 0$.

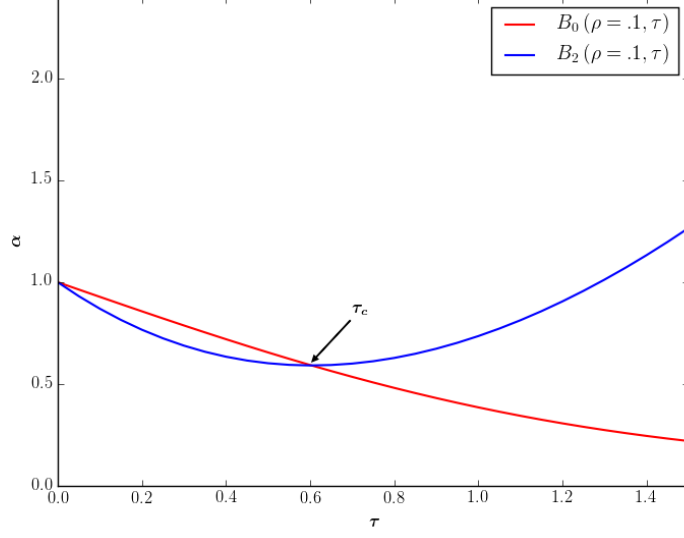


FIG. 5: The transition boundary is where the red and blue curves meet at the critical τ_c . Unlike ℓ_1 -norm minimization, the slope at this point is not zero and there exists an additional linear term to the B_2 at the critical τ_c for Elastic Net. In the text we will see that this results in different critical behavior for Elastic Net.

For the corresponding single variable optimization problem, we can still use the soft-thresholding function described in Eq. (26). The estimated value of \hat{x} is once more given by $\eta_{\text{soft}}(t; \theta)$, but with $t = \frac{x_0 + \xi}{1 + \lambda_2 \sigma_{\text{eff}}^2}$ and $\theta = \frac{\lambda_1 \sigma_{\text{eff}}^2}{1 + \lambda_2 \sigma_{\text{eff}}^2}$.

As before, we start in the perfect reconstruction phase, where $\sigma^2, \sigma_{\text{eff}}^2, \sigma_{\xi}^2 \rightarrow 0$ with $\tau = \frac{\lambda \sigma_{\text{eff}}^2}{\sigma_{\xi}^2}$ of order one. In this phase we ignore the case of non-zero x_0 leading to $\hat{x} = 0$. The contribution to MSE for the non-zero x_0 is slightly different

$$\begin{aligned} \rho[(\hat{x} - x_0)^2]_{x_0, \xi}^{\text{av}} &= \rho \left[\left(\frac{x_0 + \xi - \lambda_1 \sigma_{\text{eff}}^2 \text{sgn}(\hat{x})}{1 + \lambda_2 \sigma_{\text{eff}}^2} - x_0 \right)^2 \right]_{x_0, \xi}^{\text{av}} \\ &\approx \frac{\rho}{(1 + \lambda_2 \sigma_{\text{eff}}^2)^2} \left\{ \sigma_{\xi}^2 + (\lambda_1 \sigma_{\text{eff}}^2)^2 \left(1 + \frac{\lambda_2^2}{\lambda_1^2} [x_0^2]_{x_0}^{\text{av}} + \frac{\lambda_2}{\lambda_1} [|x_0|]_{x_0}^{\text{av}} \right) \right\} \end{aligned} \quad (60)$$

The key approximation is that $[x_0 \text{sgn}(\hat{x})]_{x_0, \xi}^{\text{av}} \approx [|x_0|]_{x_0}^{\text{av}}$, since in this limit typically $|\xi| \ll |x_0|$ implying \hat{x} and x_0 have the same sign. The other source of error is the event when the x_0 is zero but \hat{x} has fallen outside the interval $[-\theta, \theta]$ and has been estimated to be non-zero. In this case, the contribution to MSE is

$$\begin{aligned} (1 - \rho)[\hat{x}^2]_{x_0, \xi}^{\text{av}} &= 2(1 - \rho) \int_{\lambda_1 \sigma_{\text{eff}}^2}^{\infty} \frac{d\xi}{\sqrt{2\pi\sigma_{\xi}^2}} e^{-\frac{\xi^2}{2\sigma_{\xi}^2}} \left(\frac{\xi - \lambda_1 \sigma_{\text{eff}}^2}{1 + \lambda_2 \sigma_{\text{eff}}^2} \right)^2 \\ &= \frac{2\sigma_{\xi}^2(1 - \rho)}{(1 + \lambda_2 \sigma_{\text{eff}}^2)^2} \{ (1 + \tau^2)\Phi(\tau) - \tau\phi(\tau) \}. \end{aligned} \quad (61)$$

Combining Eq. (60) and (61) in the self-consistency equation for σ_{ξ}^2 and remembering that $\sigma_{\xi}^2, \sigma_{\text{eff}}^2 \rightarrow 0$ with $\tau = \frac{\lambda_1 \sigma_{\text{eff}}^2}{\sigma_{\xi}^2}$ order one, we have

$$\begin{aligned} \alpha &= 2(1 - \rho) \{ (1 + \tau^2)\Phi(\tau) - \tau\phi(\tau) \} \\ &\quad + \rho \left\{ 1 + \tau^2 \left(1 + \frac{\lambda_2^2}{\lambda_1^2} [x_0^2]_{x_0}^{\text{av}} + \frac{\lambda_2}{\lambda_1} [|x_0|]_{x_0}^{\text{av}} \right) \right\}. \end{aligned} \quad (62)$$

The equation for $\hat{\rho}$ remains the same in this limit. The denominator $1 + \lambda_2 \sigma_{\text{eff}}^2$ does not matter for the thresholding condition. As a result once more

$$\hat{\rho} = 2(1 - \rho)\Phi(\tau) + \rho. \quad (63)$$

TABLE III: Elastic Net

$\pi(x) \sim Fx^\gamma$ with $\gamma > -1$	
Input Variables	Scaling Functions
$\lambda \rightarrow 0, \sigma_\zeta^2 = 0$	$MSE \sim (\alpha_c - \alpha)^{2/(1+\gamma)}$
$\lambda \rightarrow 0, \sigma_\zeta^2 \neq 0$	$MSE \sim (\sigma_\zeta^2)^{2/(3+\gamma)}$
$\lambda \neq 0, \sigma_\zeta^2 = 0$	$MSE \sim \lambda^{2/(2+\gamma)}$
$\lambda \neq 0, \sigma_\zeta^2 \neq 0$	$MSE \sim \lambda^{2/(2+\gamma)}$ $\&\zeta \sim (\sigma_\zeta^2)^{2/(3+\gamma)}$
$\pi(x) = 0$ for $ x < \Delta$	
$\lambda \rightarrow 0, \sigma_\zeta^2 = 0$	$MSE \sim \frac{1}{\ln(1/(\alpha_c - \alpha))}$
$\lambda \rightarrow 0, \sigma_\zeta^2 \neq 0$	$MSE \sim \frac{1}{\ln(1/\sigma_\zeta^2)}$
$\lambda \neq 0, \sigma_\zeta^2 = 0$	$MSE \sim \frac{1}{\ln(1/\lambda)}$
$\lambda \neq 0, \sigma_\zeta^2 \neq 0$	$MSE \sim \frac{1}{\ln(1/\sigma_\zeta^2)}$ $\&\zeta \sim \frac{1}{\ln(1/\lambda)}$

Table presents critical exponents for Elastic Net near phase transition in terms of relevant input parameters, α , λ , and σ_ζ^2 . In order to relate with compressed-sensing literature, we have set $\sigma^2 = 1$, i.e. $\vartheta = \lambda$.

On the other hand, the condition for the phase boundary is $\alpha = \hat{\rho}$. Thus, for the Elastic Net method, the phase boundary is obtained by solving and eliminating τ from

$$\begin{aligned} \alpha &= 2(1 - \rho)\{(1 + \tau^2)\Phi(\tau) - \tau\phi(\tau)\} \\ &\quad + \rho\left\{1 + \tau^2\left(1 + \frac{\lambda_2^2}{\lambda_1^2}[x_0^{\text{av}}]_{x_0} + \frac{\lambda_2}{\lambda_1}[|x_0|]_{x_0}^{\text{av}}\right)\right\} \end{aligned} \quad (64)$$

$$\alpha = 2(1 - \rho)\Phi(\tau) + \rho \quad (65)$$

In the case of Gaussian $\pi(x_0)$ with variance $\sigma_{x_0}^2$, the key dimensionless parameter is $\frac{\lambda_2 \sigma_{x_0}}{\lambda_1}$, which determines the relative strength of the quadratic penalty term. It is important to note that unlike the ℓ_1 -norm minimization, the relation $\frac{dA_2}{d\tau} = 2\frac{A_2 - A_0}{\tau}$ in remark 4 does not hold for Elastic Net. Thus, Taylor expansion of Eq. (64) (equivalent to the A_2 term in Eq. (40)) near the transition point has linear contribution with positive slope as well as quadratic one (See Fig. 5). The theoretical critical exponents can be derived in the same way as described in section V. We only mention the result in the table III.

VI. NUMERICAL EXPERIMENTS

This section describes the numerical implementation for examining critical exponents that we obtained in section V and comparison with the numerical result. First, we compute MSE for ℓ_1 -norm minimization and Elastic Net (see Fig. 6). The matrix \mathbf{H} is obtained by first filling it with independent samples of a Gaussian distribution with variance $1/M$.

In this example, $N = 200$, $K = 30$, the original signal \mathbf{x} contains 30 randomly placed elements driven from a standard Gaussian distribution, i.e. $\gamma = 0$. The numerical experiment is carried out using CVXOPT quadratic programming [1] and for $\lambda_1 = 1\text{E} - 8$ and $\lambda_2 = 0, .4, .8$ of λ_1 (to relate with compressed-sensing literature, we have set $\sigma^2 = 1$, i.e. $\vartheta = \lambda$). As it can be seen, the reconstruction error exhibits a higher mean squared error (MSE) with respect to the theoretical result.

Next, we confirm the exponent in Eq. (52) by plotting the theoretical expression in Eq. (38). This is shown in Fig. 7.

In the end, we consider the important case where the external noise is non-zero and we are looking for a trade-off for λ where the reconstruction error minimizes. Once more, using Eq. (38), this is shown in Fig. 8. One can see that near the transition, having a non-zero λ does not help with the error on recovery of the signal. However, further from the transition line, the non-zero value of λ can result in lower MSE.

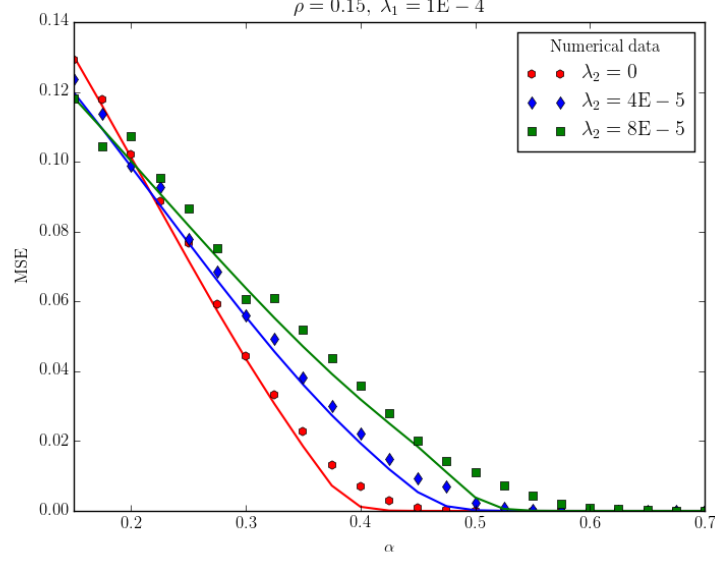


FIG. 6: Comparison of MSE for different λ_2 . Each solid curve represents the theoretical estimate for MSE as described in Sec. IV B and Sec. V D. Numerical data for different λ_2 is shown with the markers. We use CVXOPT quadratic programming to find MSE for 3 values of λ_2/λ_1 : 0, 0.4, 0.8. Notice that, for the Elastic Net ($\lambda_2 \neq 0$, the transition happens at higher α compared to ℓ_1 -norm minimization.

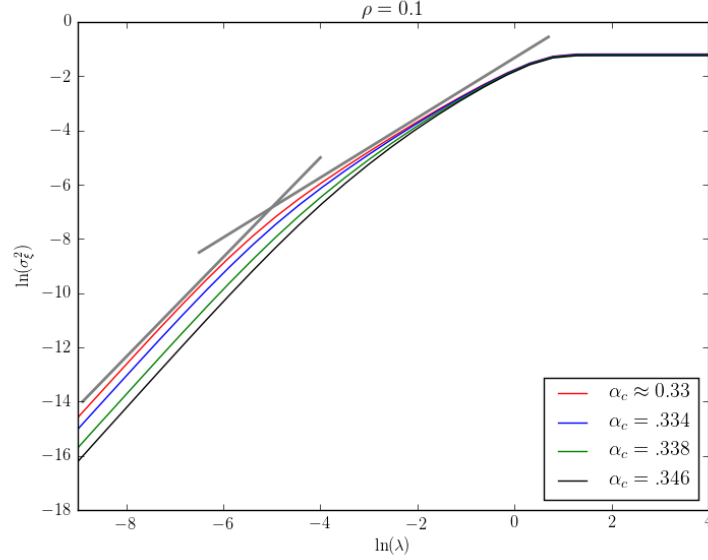


FIG. 7: Following the trends where the curves merge, we can find the critical exponent near phase transition. The grey lines show two different exponents near the transition going from the slope 1.33 to 2.

VII. CONCLUSION AND SUMMARY OF RESULTS

We presented a different approach to study the statistical properties of compressed sensing problems than the standard replica formalism and message-passing algorithms. We exploited local susceptibility obtained to treat the simple case of Ridge Regression and then to find a simple way to arrive at the two phases and the phase boundary known for the Basis Pursuit. We showed that this transition is continuous (second order) and found a variety of

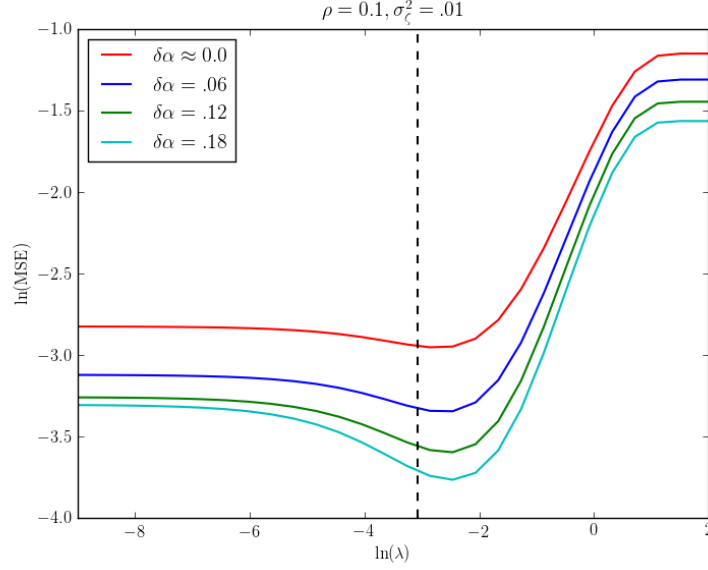


FIG. 8: Varying λ sweeps out entire optimal tradeoff curves. The vertical black dashed line is located at $\ln(\lambda) = \frac{2}{3} \ln(\sigma_\epsilon^2)$ in which the theoretical minimum error near phase transition occurs.

critical behaviors, including scaling functions and critical exponents that are uniquely determined by the universality class of the phase transition for Basis Pursuit and Elastic Net.

Throughout this paper we stressed the important role of local susceptibility introduced by the zero-temperature cavity method as a powerful tool in sparse recovery problems. It turns out that the perfect reconstruction phase corresponds to vanishing average local susceptibility, indicating that the solution of the optimization problem has an underlying robustness to perturbations in this phase. This implies that the structure of susceptibility enjoys unique properties and applicability far beyond the standard sparse setting traditionally considered in compressed sensing.

Acknowledgment

This work was supported by the National Science Foundation INSPIRE (track 1) award 1344069. The final version of this paper was written while two of the authors (AMS and MR) were visiting Simons Center for Data Analysis at Simons Foundation. We are grateful for their hospitality.

Appendix A: Ridge Regression via Singular Value Decomposition

For the sake of completeness, in this appendix, we derive Eqs. (19) and (20) in section IV A from a formal singular value decomposition point of view. Elementary derivation leads us to an explicit expression:

$$\hat{\mathbf{x}} = \frac{\mathbf{H}^T \mathbf{H}}{\sigma^2} \left[\frac{\mathbf{H}^T \mathbf{H}}{\sigma^2} + \lambda \mathbf{I}_N \right]^{-1} \mathbf{x}_0 = \sum_{i=1}^M \frac{s_i^2}{s_i^2 + \lambda \sigma^2} \mathcal{V}_i (\mathcal{V}_i^T \mathbf{x}_0). \quad (\text{A1})$$

where we use the singular vector basis of the matrix \mathbf{H} , with \mathbf{s}_i being the non-zero singular values, and \mathcal{V}_i the corresponding right singular vectors. When we take the limit of vanishing σ^2 , we just have a projection of the N dimensional vector \mathbf{x}_0 to an M -dimensional projection spanned by \mathcal{V}_i 's. In other words

$$x_a = \sum_{b=1}^N \sum_{i=1}^M \mathcal{V}_{ia} \mathcal{V}_{ib} x_{0b} = \sum_{a=1}^N P_{ab} x_{0a} \quad (\text{A2})$$

\mathbf{P} being the projection matrix. For random \mathbf{H} , \mathcal{V}_i 's are just a random choice of M orthonormal vectors. Thus, the properties of the estimate depends on the statistics of the projection matrix to a random M -dimensional subspace.

$$[P_{ab}]_{\mathbf{H}}^{\text{av}} = \sum_{i=1}^M [\mathcal{V}_{ia} \mathcal{V}_{ib}]_{\mathbf{H}}^{\text{av}} = \sum_{i=1}^M \frac{\delta_{ab}}{N} = \alpha \delta_{ab} \implies [\hat{x}_a]_{\mathbf{H}}^{\text{av}} = \alpha x_{0a} \quad (\text{A3})$$

For variance, we need to think of second order moments of the matrix elements of \mathbf{P} , particularly, $[P_{ab}P_{ac}]_{\mathbf{H}}^{\text{av}}$. We could parametrize $[P_{ab}P_{ac}]_{\mathbf{H}}^{\text{av}} = A\delta_{bc} + B\delta_{ab}\delta_{bc}$. Since \mathbf{P} is a projection operator, $\mathbf{P}^2 = \mathbf{P}$ and it is a symmetric matrix. Hence,

$$\sum_a [P_{ab}P_{ac}]_{\mathbf{H}}^{\text{av}} = \sum_a [P_{ba}P_{ac}]_{\mathbf{H}}^{\text{av}} = [P_{bc}]_{\mathbf{H}}^{\text{av}} = \alpha \delta_{bc}. \quad (\text{A4})$$

In the limit of $M, N \rightarrow 0$ with α fixed, the distribution of P_{aa} gets highly concentrated around the mean α . As a result,

$$[P_{aa}P_{aa}]_{\mathbf{H}}^{\text{av}} \approx \left(\sum_a [P_{aa}]_{\mathbf{H}}^{\text{av}} \right)^2 = \alpha^2. \quad (\text{A5})$$

Using the two constraints, represented by Eqs. (A4) and (A5), we can determine A and B , in the large M, N limit, leading to,

$$[P_{ab}P_{ac}]_{\mathbf{H}}^{\text{av}} \approx \frac{\alpha(1-\alpha)}{N} \delta_{bc} + \alpha^2 \delta_{ab}\delta_{bc}. \quad (\text{A6})$$

The variance is now given by,

$$\begin{aligned} & [\hat{x}_{0a}^2]_{\mathbf{H}}^{\text{av}} - ([\hat{x}_{0a}]_{\mathbf{H}}^{\text{av}})^2 \\ &= \sum_a \left[\frac{\alpha(1-\alpha)}{N} \delta_{bc} + \alpha^2 \delta_{ab}\delta_{bc} \right] x_{0b}x_{0c} - (\alpha x_{0a})^2 \\ &= (1-\alpha)\alpha \rho [x_0^2]_{x_0}^{\text{av}} \end{aligned} \quad (\text{A7})$$

recovering our earlier result.

-
- [1] Martin S Andersen, Joachim Dahl, and Lieven Vandenbergh. CVXOPT: A python package for convex optimization, 2010.
 - [2] Emmanuel Candès and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.
 - [3] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.
 - [4] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.
 - [5] David L Donoho. For most large underdetermined systems of linear equations the minimal. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.
 - [6] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.
 - [7] David L Donoho and Jared Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9446–9451, 2005.
 - [8] Surya Ganguli and Haim Sompolinsky. Statistical mechanics of compressed sensing. *Physical review letters*, 104(18):188701, 2010.
 - [9] Yoshiyuki Kabashima, Tadashi Wadayama, and Toshiyuki Tanaka. A typical reconstruction limit for compressed sensing based on ℓ_p -norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):L09003, 2009.
 - [10] Ryogo Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255, 1966.
 - [11] Marc Mézard, Giorgio Parisi, and MA Virasoro. Sk model: The replica solution without replicas. *Europhys. Lett*, 1(2):77–82, 1986.
 - [12] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond*, volume 9. World scientific Singapore, 1987.
 - [13] Mohammad Ramezanali, Partha P Mitra, and Anirvan M Sengupta. Cavity method for penalized regression optimization. *Preprint*, 2015.

- [14] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.
- [15] Andrey Nikolayevich Tikhonov. On the stability of inverse problems. 39(5):195–198, 1943.
- [16] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.